

## 小组周报（2014.11.10-2014.11.16）

本周工作：

本周一直在考虑温州的数据，以为要用温州数据。所以大家的工作主要还是围绕这个数据展开。

- 1) 通话数据的处理部分：这部分是想从里面找到用户的手机号码和微博账号之间的对应关系。之前的想法可能有点简单，以为根据用户的上网浏览微博网页的数据找到用户的一条 post 信息然后再抓取猜测的用户的微博进行验证就可以。后面发现实际用户的用手机 post 的一条微博在该数据中会有多条记录，并且时间范围有的差别还挺大的。并且考虑到就算用户当时用手机 post 了一条微博，到现在也可能会把这条微博删掉，仍然无法进行验证。在这上面走了一些弯路。后面想的方法的只要猜测者有一条转发的微博的转发人、转发的时间和转发方式（手机）都一样就认为是完全匹配了，认为猜测的是正确的。这部分是杨哲一直在做。
- 2) 本体构建部分：我们把之前看的 palantir 视频中出现的本体都考虑进去，希望可以构建一个相对全面一些的本体。之认为看 palantir 的视频中他们的本体不是特别大，所以目前我们仅是构建了一个比较小的本体系统。针对陈老师考虑的要我们构建一个全面的本体，后面仍然要继续进行丰富完善。这部分由王琦在做。
- 3) 底层的数据整合的数据源的处理：目前粗糙地完成了 csv 数据集和数据库数据集的简单整合问题。数据库的整合采用 d2rq, d2rq 完全支持了 mysql、oracle 等数据库，可以自动分析数据库的关联结构、查询等问题。csv 数据集需要用户自己手动制定每一列的含义，与 global ontology 的对应关系。这部分由李嘉华和马晓红做。

下周工作：

1. 周一至周三调研 1) 数据整合中不同数据源的可信度问题  
2) 如果处理数据整合中数据的不一致和模糊性问题。
2. 之前有看到 google 的知识图谱 freebase，建立的本体结构都是跟人相关的，

并且相对比较全，不过目前我们没有办法把 **freebase** 上的结构爬下来。本体的构建这部分王琦会一直继续做的。